

dPABBs: A Novel *in silico* Approach for Predicting and Designing Anti-biofilm Peptides

Pooja Gupta¹, Rakesh Kumar^{1#}, Arun Sharma^{1#*}, Anshu Bhardwaj^{1#*}

¹Open Source Drug Discovery (OSDD) Unit, Council of Scientific and Industrial Research, New Delhi, India.

#Academy of Scientific and Innovative Research, Delhi, India.

*bioinfo.arun@gmail.com

*anshub@osdd.net

*To whom the correspondence should be addressed:

Anshu Bhardwaj

Scientist, Open Source Drug Discovery (OSDD) Unit

Faculty Coordinator & Assistant Professor,

The Academy of Scientific & Innovative Research (AcSIR)

Council of Scientific and Industrial Research (CSIR)

Anusandhan Bhawan, 2 Rafi Marg, Delhi, India

+ 91 011 23470121

Skype: anshu.bhardwaj

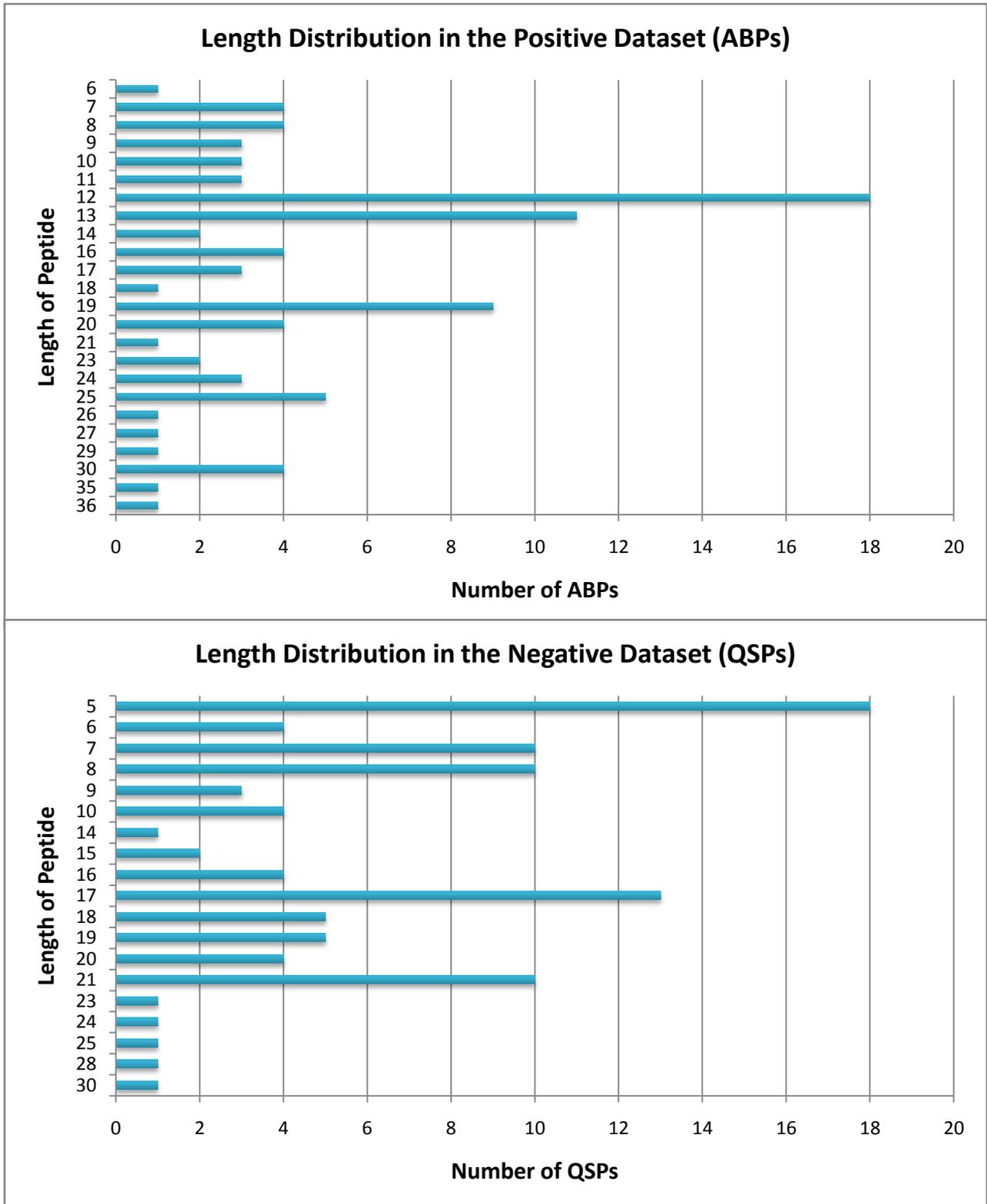
<http://ab-openlab.csir.res.in/>

[A] List of the ABPs tested against the biofilms of Gram positive and Gram negative bacteria

S. No.	GRAM STATUS	BACTERIA	NUMBER OF ACTIVE PEPTIDES
1	Gram negative	<i>Acinetobacter baumannii</i>	3
2	Gram negative	<i>Burkholderia ambifaria</i>	1
3	Gram negative	<i>Burkholderia cenocepacia</i>	4
4	Gram negative	<i>Burkholderia dolosa</i>	1
5	Gram negative	<i>Burkholderia multivorans</i>	1
6	Gram negative	<i>Burkholderia pseudomallei</i>	3
7	Gram negative	<i>Burkholderia stabilis</i>	1
8	Gram negative	<i>Escherichia coli</i>	2
9	Gram negative	<i>Escherichia coli</i> ATCC 25922	4
10	Gram negative	<i>Escherichia coli</i> HM22	2
11	Gram negative	<i>Escherichia coli</i> O-157 / ATCC 43895	2
12	Gram negative	<i>Escherichia coli</i> RP437	2
13	Gram negative	<i>Klebsiella pneumoniae</i> ATCC 13883	1
14	Gram negative	<i>Porphyromonas gingivalis</i> ATCC 33277	1
15	Gram negative	<i>Pseudomonas aeruginosa</i>	18
16	Gram negative	<i>Pseudomonas aeruginosa</i> ATCC 15442	1
17	Gram negative	<i>Pseudomonas aeruginosa</i> ATCC 27853	10
18	Gram negative	<i>Pseudomonas aeruginosa</i> PA14	7
19	Gram negative	<i>Pseudomonas aeruginosa</i> PAO1 / ATCC 15692	13
20	Gram negative	<i>Salmonella enterica</i> Typhimurium ATCC14028	1
21	Gram positive	<i>Bacillus subtilis</i>	1
22	Gram positive	<i>Enterococcus faecalis</i>	2
23	Gram positive	<i>Enterococcus faecium</i> ATCC 19434	2
24	Gram positive	<i>Lactobacillus salivarius</i>	1
25	Gram positive	<i>Listeria iavanovii</i> Li4pVS2	1
26	Gram positive	<i>Propionibacterium acnes</i> ATCC 6919	1
27	Gram positive	<i>Staphylococcus aureus</i>	11
28	Gram positive	<i>Staphylococcus aureus</i> ATCC 25923	4
29	Gram positive	<i>Staphylococcus aureus</i> ATCC 29213	2
30	Gram positive	<i>Staphylococcus aureus</i> ATCC 29737	2
31	Gram positive	<i>Staphylococcus aureus</i> ATCC 43300	2
32	Gram positive	<i>Staphylococcus aureus</i> ATCC 6538	1
33	Gram positive	<i>Staphylococcus aureus</i> Smith diffuse	2
34	Gram positive	<i>Staphylococcus epidermidis</i> ATCC 35984	2
35	Gram positive	<i>Streptococcus gordonii</i>	1
36	Gram positive	<i>Streptococcus mutans</i>	4
37	Gram positive	<i>Streptococcus mutans</i> ATCC 25175	1

38	Gram positive	<i>Streptococcus mutans</i> Clarke UA159 / ATCC 700610	9
39	Gram positive	<i>Streptococcus mutans</i> KCTC 3065	1
40	Gram positive	<i>Streptococcus oralis</i> ATCC 10557	1
41	Gram positive	<i>Streptococcus parasanguinis</i> ATCC 903	1
42	Gram positive	<i>Streptococcus salivarius</i>	1
43	Gram positive	<i>Streptococcus salivarius</i> ATCC 7073	1
44	Gram positive	<i>Streptococcus sanguinis</i>	1
45	Gram positive	<i>Streptococcus sanguinis</i> ATCC 10556	1
46	Gram positive	<i>Streptococcus sanguinis</i> NY101	1
47	Gram positive	<i>Streptococcus sobrinus</i> ATCC 6715	1

[B] Length distribution of the ABPs and QSPs (Frequency of peptides of a particular length)



[C] GRAVY scores for ABPs and QSPs

ABPs (80)	GRAVY SCORE	QSPs (88)	GRAVY SCORE
FFGWLIKGAIHAGKAIHGLIHRRRH	0.00	AKDEH	-2.46
FIKHFIHRFGGGFKKFWKWFRRF	-0.43	AKTVQ	-0.42
FIKHFIHRFGGGRWRWRWF	-0.63	ARNQT	-2.08
FIKHFIHRFSATRWRWRWF	-0.55	DRVGA	-0.48
FKCRRWQWRMCKLG	-1.42	EKMIG	-0.28
FLGALFKALSLL	1.48	ERGMT	-1.44
FLSLIPHIVSGVASIAKHF	1.26	ERNNT	-3.14
FRIRVRV	0.31	ERPVG	-1.16
GIFSKLAGKKIKNLLISGLKG	0.34	GKAEF	-0.64
GIGAVLKVLTTGLPALISWIKRKRQQ	0.27	LPFEF	0.86
GIGLFLHSAGLFGALFVGEIMKS	1.22	LPFEH	-0.34
GKIIKLGKASLL	0.71	MKAEH	-1.38
GLFDVIKKVASVIGGL	1.28	MPFEF	0.48
GRRRSVQWCA	-1.37	QRGMI	-0.40
GRWKRWRKKWKKLWKKLS	-2.11	SKDYN	-2.60
GWGSFFKAAHVKGKVGKAALTHYL	-0.07	SRNVT	-1.06
GWKKWLRKGAKHLGQAAIK	-0.85	VPFEF	0.94
HLGHHALDHLK	-0.30	ADLPFEF	0.37
IGKEFKRIVERIKRFLRELVRPLR	-0.51	AITLFI	3.03
IGKEFKRIVQRIK	-0.58	LVTLVTV	3.19
IGKEFKRIVQRIKDFLRNL	-0.46	NEVPFEF	-0.33
ILPWKWPWWPWRR	-1.07	SDLPFEH	-0.86
IRIKIRIK	0.15	SDMPFEF	-0.27
IRVKIRVKIRVK	0.08	SYPGWSW	-0.96
IRWRIRVWVRRIC	0.01	GLDWWSL	0.16
IVWKIKRWVWGR	-0.21	ALILTVS	2.55
KIFGAIWPLALGALKNLIK	0.97	FLVMFLSG	2.26
KKHRKHRKHRKHGGSGGSKNLRRIIRKGIHI IKKYG	-1.68	LFSVLVLAG	2.38
KQFRIRVRV	-0.58	LFVTVLVG	2.74
KRFRIRVRVIRK	-0.84	VAVLVVLAG	2.93
KRIVQRIKDFLRNLVPRTES	-0.76	DIIVVGG	2.24
KRWRIRVRVIRKC	-0.87	DILIVVGG	2.15
KRWRWIVRNIRR	-1.54	EGIIVVVG	2.42
KRWWKWWRRRC	-2.24	ILPYFAGCL	1.77
KSKEKIGKEFKRIVQRIKDFLRNLVPRTES	-1.05	CVFSLFKKCN	0.65
KWCFRVCYRGICYRKC	-0.48	ILSGAPCIPW	1.18
KWFWKFKVFKVK	-0.05	AILPYFAGCL	1.77
KWKLKFKIGIGKFLHSAKFF	-0.31	DLRNIFLKIKFKKK	-0.63
KWKVFKKIEKMGRRNIRNGIVKAGPAIAVLG EAKAL	-0.07	DLRGVNPWGWIFGR	-0.46

LAHQKPFIRKSYKCLHKRCR	-1.05	SNLVECVFSLFKKCN	0.45
LAREYKKIVEKLRWLRQVLRTLR	-0.74	EMRKSNNNFFHFLRRI	-1.05
LLGDFFRKSKEKIGKEFKR	-1.06	DSRIRMGDFSKLFGK	-0.48
LLGDFFRKSKEKIGKEFKRIVQRIK	-0.75	EIRQTHNIFNFFKRR	-0.94
LRIRWIFKR	-0.30	EMRKP DGALFNLFRRR	-1.09
NEEGFFSARGHRPLDGGGKKKKKK	-1.78	DKRLPYFFKHLFSNRTK	-1.14
RFKRVARVIW	-0.08	DRRDPRGIIGIGKKLFG	-0.67
RFRRLFRRIRVRVLKKI	-0.26	DWRISSETIRNLIFPRRK	-0.96
RFWKVRVKYIRF	-0.42	EMRISRIILDFLFLRKK	0.19
RIVQRIKDFLRNLVPRTES	-0.59	EMRLPKILRDFIFPRKK	-0.66
RIWVIWRR	-0.26	EMRLSKFFRDFILQRKK	-0.83
RKSKEKIGKEFKRIVQRIK	-1.48	ESRLPKILLDFLFLRKK	-0.06
RKSKEKIGKEFKRIVQRIKDFLRNL	-1.17	ESRLPKIRDFIFPRKK	-0.88
RKSYKALHKRAR	-1.93	ESRIDILLDFLQRRK	-0.38
RKSYKCLHKRCR	-1.81	SGSLSTFFRFLNRSFTQ	-0.13
RPAFRKAAFRVMRACV	0.13	STFFRFLNRSFTQALGK	-0.16
RRWIRVAVILRV	0.69	GKATSSISKCVSFFKCC	0.22
RRWVWRIVQRR	-0.89	LSTFFRFLNRSFTQALGK	0.06
RWRWRW	-2.70	SGSLSTFFRFLNRSFTQA	-0.02
RWRWRWF	-1.91	SGTLSTFFRFLNRSFTQA	-0.02
RWRWRWRW	-2.70	SLSTFFRFLNRSFTQALG	0.64
TFRFLNRGGGKNLRIIRKGIHIKKY	-0.33	DIRHRINNSIWRDIFLKRK	-1.06
TFRFLNRGGGWGSFFKAAHVGKL	-0.10	GKPASNLVECVFSLFKKCN	0.14
TLISWIKNKRKQCRPRVSRRRRRRGRRRR	-2.11	SGSLSTFFRFLNRSFTQAL	0.18
TLISWIKNKRKQRPRVSRRRRRRGRRRR	-2.27	SLSTFFRFLNRSFTQALGK	0.02
TLISWIKNKRKQRPRVSRRRRRRGRRRRC	-2.11	GSLSTFFRFLNRSFTQALGK	-0.01
VQFRIRVRIVIRK	0.27	SGSLSTFFRFLNRSFTQAGK	-0.24
VQWRIRVRVIKK	-0.34	SGSLSTFFRFLNRSFTQALG	0.15
VRLIRAVRAWRV	0.47	SGSLSTFFRFLNRSFTQALK	-0.03
VRLIVAVRIWRR	0.69	SGSLSTQFRLNRSFTQALGK	-0.34
VRLIVRIWRR	0.23	SGSLSTFFLLNRSFTQALGK	0.35
VRLIWA VRIWRR	0.27	SGSLSTFFRFLRSFTQALGK	0.30
VRLRIRVA	0.63	SGSLSTFFRFLNASFTQALGK	0.26
VRLRIRWWVLRK	-0.27	SGSLSTFFRFLNRSFTQALGA	0.23
VRWARVARILRV	0.47	SGSLSTFFRFLNRSFTQALGK	-0.04
VTCDVLSFEAKGIAVNH	0.61	SGSLSTFFRFLNRSFTQALGV	0.34
WIVVIWRRKRRC	-0.64	SGSLSTFFRFLNRSQTQALGK	-0.34
WKLLSKAQEKFGKNKSR	-1.48	SGSLSTFFRLQNRSFTQALGK	-0.34
WRWRVRVWR	-1.37	SQKGVYASQRSFVPSWFRKIFRN	-0.70
YAPWTNA	-0.63	AGTKPQGKPASSISKCVSFFKCC	-0.22
YAPWTNF	-0.49	SINSIQGKATSNLVECVFSLFKKCN	0.16
Average ABPs	-0.53	EQLSFTSIGILQLLTIGTRSCWFFYCRY	0.49
		SGWMDYINGFLKGGGQRTLPTKDY NIPQV	-0.53
		SRNAT	-1.54

		AIFILAS	2.63
		SIFTLVA	2.23
		EIIIVGG	2.24
		ESRVSRIILDFLFQRKK	-0.41
		SGSLSTFFRLFNFSFTQALGK	0.30
		Average QSPs	0.10

For all other properties (such as amino acid percentage composition, physicochemical properties) similar analysis (two sample t-test assuming unequal variance, using Microsoft Office 2007) was performed to determine statistical significance (p-value < 0.05).

[D] Prediction on FDA approved biotech peptides and examples of active/inactive mutant peptides with improved SVM scores

DrugBank ID	Parent peptide	Predicted as Biofilm-active (No. of models)	Mutated peptide	Predicted as Biofilm-active (by SVM Whole AAC)	SVM score (Parent / Mutant)
DB04921	DENPVVHFFKNIVTPRT	Yes (1)	DENPVVH R FKNIVTPRT	Yes	0.46 / 0.92
DB01309	FVKQHLCGSHLVEALYLVCGERGFFYTPET	Yes (3)	FVKQHLCGSHLVEALYLVCGERGFFY T HET	Yes	0.31 / 0.71
DB01306	FVNQHLCGSHLVEALYLVCGERGFFYTDKT	Yes (3)	FVNQHLCG H HLVEALYLVCGERGFFYTDKT	Yes	0.33 / 0.72
DB00046	FVNQHLCGSHLVEALYLVCGERGFFYTKPT	Yes (3)	FVNQHLCGSHLVEALYLVCGERGFFY T K H T	Yes	0.30 / 0.71
DB00030; DB00071; DB08914	FVNQHLCGSHLVEALYLVCGERGFFYTPKT	Yes (3)	FVNQHLCGSHLVEALYLVCGERGFFY T H K T	Yes	0.30 / 0.71
DB00047	FVNQHLCGSHLVEALYLVCGERGFFYTPKTRR	Yes (3)	FVNQHLCG K HLVEALYLVCGERGFFYTPKTRR	Yes	0.56 / 0.96
DB05875	RPKPQQFFGLM	Yes (3)	R H K PKPQQFFGLM	No	-1.31 / -0.54
DB08888	APSFDCGKPQVEPKKCPGR	Yes (4)	A H S FDGKPKQVEPKKCPGR	No	-1.12 / -0.56
DB01307	FVNQHLCGSHLVEALYLVCGERGFFYTPK	Yes (4)	FVNQHLCGSHLVEALYLVCGERGFFY T H K	Yes	0.29 / 0.73
DB06655	HAEGTFTSDVSSYLEGQAAKEEFIIAWLVKGRG	Yes (4)	HAEGTFT S H V SSYLEGQAAKEEFIIAWLVKGRG	Yes	0.42 / 0.83
DB05332	IEGPTLRQWLAARA	Yes (4)	IEGPTLRQW L VARA	Yes	0.13 / 0.79
DB05685	MKHLKLLKCVFLVKSQGVNDNEEGFFS	Yes (4)	MKHLKLLKCVFLV K H Q GVNDNEEGFFS	Yes	-0.37 / 0.10
DB04897	KLLLLKLLKLLKLLKLLK	Yes (5)	K ALLKLLKLLKLLKLLK	No	-0.49 / -0.33
DB06285	SVSEIQLMHNLGKHLNSMERVEWLRKKLQDVHNF	Yes (3)	V VSEIQLMHNLGKHLNSMERVEWLRKKLQDVHNF	Yes	0.57 / 0.97
DB01284	SYSMEHFRWGKPVGKKRRPVKVYP	Yes (4)	SYSMEHFRWG K H V GKKRRPVKVYP	Yes	0.57 / 1.06
DB00017	CSNLSTCVLGKLSQELHKLQTYPRNTGSGTP	Yes (2)	C H N LSTCVLGKLSQELHKLQTYPRNTGSGTP	No	-0.69 / -0.30
DB00067	CYFQNCPRG	Yes (2)	CYFQNC R RG	No	-1.21 / -0.53
DB00107	CYIQNCPLG	Yes (2)	CYIQNC H LG	No	-1.24 / -0.47
DB00006	FPRPGGGNGDFEEIPEEYL	Yes (2)	F H R PGGGNGDFEEIPEEYL	No	-1.63 / -1.23
DB00047	GIVEQCCTSICSLYQLENYCG	Yes (2)	GIVEQC C T H ICSLYQLENYCG	No	-0.89 / -0.43

DrugBank ID	Parent peptide	Predicted as Biofilm-active (No. of models)	Mutated peptide	Predicted as Biofilm-active (by SVM Whole AAC)	SVM score (Parent / Mutant)
DB00030; DB00046; DB00071; DB01306; DB01307; DB01307; DB01309; DB08914	GIVEQCCTSICSLYQLENYCN	Yes (2)	GIVEQCCTHICSLYQLENYCN	No	-0.95 / -0.48
DB08900	HGDGSFSDEMNTILDNLAARDFINWLIQTKITD	Yes (1)	HGVGSFSDEMNTILDNLAARDFINWLIQTKITD	No	-1.19 / -0.71
DB00021	HSDGTFTSELSRLRDSARLQRLQGLV	Yes (1)	HHDGTFTSELSRLRDSARLQRLQGLV	No	-0.93 / -0.40
DB00040	HSQGTFTSDYSKYLDSRRAQDFVQWLMNT	Yes (1)	HHQGTFTSDYSKYLDSRRAQDFVQWLMNT	No	-1.36 / -0.85
DB04985	PGDSTRKCMDLKGNK	Yes (1)	PGHSTRKCMDLKGNK	No	-0.95 / -0.15
DB00007	PHWSYLLR	Yes (2)	PHWKYLLR	Yes	-0.45 / 0.39
DB04900	SDAAVDTSEITTKDLKEKKEVVEEAEN	No	VDAAVDTSEITTKDLKEKKEVVEEAEN	Yes	-0.20 / 0.08
DB04899	SPKMVQSGSGCFGRKMDRISSSSGLGCKVLRH	No	HPKMVQSGSGCFGRKMDRISSSSGLGCKVLRH	No	-0.62 / -0.13
DB05671	TDLQERGDNDISPFSGDGPFPKD	Yes (2)	TDLQERGDNDIVPFSGDGPFPKD	No	-2.20 / -1.78
DB00010	YADAIFTNSYRKVLGQLSARKLLQDIMSQR	Yes (2)	YADAIFTNHYSRKVLGQLSARKLLQDIMSQR	No	-0.61 / -0.04
DB00109	YTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWF	Yes (2)	YTHLIHSLIEESQNQQEKNEQELLELDKWASLWNWF	No	-0.61 / -0.30

[E] Prediction on some antimicrobial peptides in clinical development. The table lists the peptide name, peptide sequence, trial phase and prediction on the peptide

Peptide Name	Peptide Sequence	Clinical Stage (Phase)	Prediction (No. of models)
Pexiganan acetate (MSI 78)	GIGKFLKKAKKFGKAFVKILKK	III	Biofilm-active (6)
Omiganan (MX-226/MBI-226)	ILRWPWWPWRRK	III	Biofilm-active (6)
Omiganan (CLS001)	ILRWPWWPWRRK	II / III	Biofilm-active (6)
Iseganan (IB-367)	RGGLCYCRGRFCVCVGR	III	Biofilm-active (5)
hLF1-11	GRRRRSVQWCA	I / II	Biofilm-active (6)
PAC-113	AKRHHGYKRKFH	IIb	Biofilm-active (6)
CZEN-002	CKPVCKPV	IIb	Biofilm-active (3)
IMX942	KSRIVPAIPVSLI	Ia	Biofilm-active (3)
OP-145	IGKEFKRIVERIKRFLRELVRPLR	II (completed)	Biofilm-active (6)
Ghrelin	GSSFLSPEHQRVQQRKESKKPPAKLQPR	II	Biofilm-active (1)

[F] Analysis of 26 QSPs with anti-biofilm activity: percentage of correct predictions made by the 6 models on dPABBs

Models	Number of correct predictions	Total number of peptides	Percentage of correct predictions (%)
SVM (20)	6	26	23.07
SVM (14)	2	26	7.69
Weka (20)	12	26	46.15
Weka (8)	14	26	53.84
SVM (NT5)	20	26	76.92
Weka (NT5)	23	26	88.46

[D`] Analyses to ascertain whether the 31 FDA-approved peptide sequences are derivatives of the 80 anti-biofilm peptides (positive dataset used in training)

In the first analysis, a database of the anti-biofilm peptides (ABPs) was created and BLAST (blastp) was performed for each of the FDA-approved peptides against this database (e-value 0.0001, other default parameters). None of the FDA-approved peptides showed a match with any ABP. From this sequence alignment data, it can be concluded that the ABPs and the FDA-approved peptides do not share any significant sequence similarity.

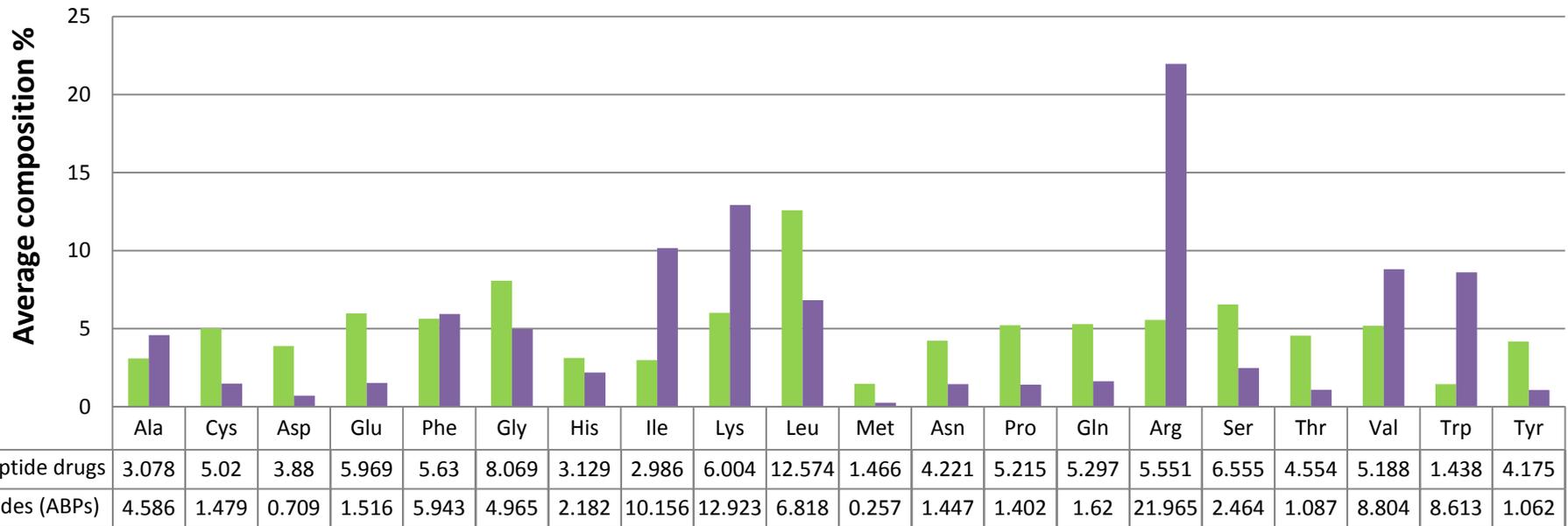
In the second analysis, a library of four amino acid-long peptides was created from each FDA-approved peptide and ABPs, using a step size of one amino acid. This library obtained from FDA-approved peptides was then compared with that obtained from the ABPs for exactly matching four-amino acid sequences. Seven four-amino acid stretches (GQAA, NEEG, EEGF, EGFF, GFFS, LKLL, WLRK) were found to be common among these. They belong to the following four FDA-approved peptides from the 31 used in the study:

FOUR-RESIDUE MOTIF	FDA-APPROVED SEQUENCE	ANTI-BIOFILM PEPTIDE
GQAA	HAEGTFTSDVSSYLE GQAA KEEFIIAWLVKGRG	GWKKWLRKGAKHL GQAA IK
NEEG	MKHLLLLLLCVFLVKSQGVND NEEG FSS	NEEG FSSARGHRPLDGGGKKKKKK
EEGF	MKHLLLLLLCVFLVKSQGVND EEGF S	NEEG FSSARGHRPLDGGGKKKKKK
EGFF	MKHLLLLLLCVFLVKSQGVND EEGF S	NEEG FSSARGHRPLDGGGKKKKKK
GFFS	MKHLLLLLLCVFLVKSQGVND EEGF S	NEEG FSSARGHRPLDGGGKKKKKK
LKLL	KLLL LKLL LLLLLLLLL	GKIKLKAS LKLL
WLRK	SVSEIQLMHNLGKHLNSMERVE WLRK KLQDVHNF	GWKK WLRK GAKHLGQAAIK

Additionally, the average composition % for each of the amino acid residue was also compared for the two sets of peptides, as shown in the graphs below. It is thus evident that the FDA-approved peptides and the ABPs have different amino acids residue compositions.

Based on these three analyses, it can be safely concluded that the 31 FDA-approved peptides and the 80 ABPs have different sequences and compositions of amino acids, and related or derivatives of each other.

[A] Average Amino Acid Residue Composition % - ABPs versus FDA-approved Peptide Sequences



[B] Fold-change in the Average Amino Acid Residue Composition % - ABPs versus FDA-approved Peptide Sequences

